

## **Collaborative Trials: Their Use and Abuse**

ERIC C. WOOD

*Lincolne Sutton and Wood, Clarence House, Clarence Road, Norwich  
NRI 1HG*

The design and operation of Collaborative Trials are discussed, with special reference to their value to Public Analysts and to the responsibilities of both the organisers and the participating laboratories.

Many Public Analysts have recently been involved in several<sup>1,2,3,4,5,6</sup> collaborative trials (referred to hereafter as Trials) of analytical methods for the determination of a component or components of foodstuffs and the like. Some Public Analysts have questioned the value of these Trials, requiring as they do the expenditure of much time by assistants whose primary function is the analysis of samples submitted officially by the Local Authority; this criticism can be answered only by consideration of the cost-effectiveness of these Trials. It may therefore be of value to discuss Trials in general from first principles, including the reasons for performing them; their value to analysts and in particular to Public Analysts; and the conditions to be observed by both organisers and participants if the results are to be of maximum benefit. It is assumed in what follows that the reader is acquainted with the fundamental principles of statistics and the meaning of such terms as "accuracy", "precision", and "standard deviation".

### **The Rationale of Collaborative Trials**

Suppose that a new law requires a minimum content of a certain component C in a foodstuff F. Public Analysts now have the duty of determining the proportion of C in any sample of F officially submitted to them and of reporting the result. Suppose also that a method for doing this has already been published but until the new law came into force had received little attention. The actions that probably would (and certainly should) be taken by a Public Analyst are as follows.

1. The method is tested on a random sample of F, (a) to see if it "works", (i.e. it reveals no obvious practical snags); (b) to ascertain whether it can be completed in a reasonable time; and (c) to ensure that it gives a well-defined quantitative result that can be satisfactorily replicated. In other words, the method must pass tests for practicability and precision.

2. Unless the sample of F tested has a known C content (often this is not so in practice), he cannot tell whether the method is also accurate; the Analyst must then make "recovery tests" after addition of known amounts of pure C to the sample.

3. Even if the method passes this test for accuracy also, this is true only for the one laboratory; two or more operators may have analysed the sample with satisfactory agreement between their results, but it is still possible that all the results were affected by a constant error, such as that due to an incorrectly standardised titrant.\*

4. The Analyst then requests other Public Analysts to analyse sub-samples of the same sample by the same method and compare their results with his. Whether two or twenty-two laboratories participate, a Trial has been made.

### **The Value of Collaborative Trials**

All these steps are essential stages in the evaluation of a method intended for use by the Public Analyst if he is going to inform his Local Authority that he is now ready to receive samples of F taken officially by the Sampling Officer and to report their content of C. For this implies that in certain cases (see below) he is prepared to issue an official certificate stating that a sample is deficient in C content to an extent that he knows is likely to result in the Local Authority taking legal proceedings against the vendor. The defence may then produce a certificate from another Public Analyst who has analysed the vendor's counterpart of the same sample and has found a higher C content than did the first Public Analyst. If the difference between the two is large enough to cause the Magistrates to dismiss the case, this will mean at the very least substantial damage to the confidence of the Local Authority in their Public Analyst. It is a major virtue of a Trial that each participating Public Analyst can feel confident that he will not find himself "out on a limb" if he gives official certificates of non-genuineness relating to samples of foodstuff F analysed by the tested method.

But this is only a qualitative approach. Whenever the law prescribes a minimum value for the content of a component in a food, the Public Analyst has to fix for himself a certain figure, less than the legal minimum, to divide the deficient samples on which he gives only an informal report from those so deficient as to justify a formal adverse certificate and consequent legal proceedings. If he chooses too low a value for this partitioning (let us call it the Criterion), too many manufacturers of inferior food will go unpunished; if it is too high, he will risk, as explained above, the possibility that a prosecution might fail. He needs quantitative data to assist him in solving his dilemma.

### **Use of Statistical Analysis**

These data are obtained by the statistical examination of the numerical results of the Trial. A well-known example of historic interest arises from the determination of extraneous water in milk from its freezing-point depression. After what has been in effect a long series of Trials, the accuracy and precision of this measurement under carefully standardised conditions are known closely

\* The author has recently become aware of a Trial in which the results of the participating laboratories agreed well with each other except that those from one were all about 20 per cent. different from the rest. This was traced to an inaccurate calibration graph; the error was unsuspected by the laboratory in question until the results from the Trial as a whole showed that an investigation was necessary.

enough to enable the risk of error in certifying non-genuineness to be stated within close limits; for example, the B.S.I. recommendation\* that  $0.525^{\circ}\text{C}$ . should be taken as the smallest f.p.d. for a genuine herd milk is based on the calculation that the risk that a genuine milk will have a smaller f.p.d. is only 0.1 per cent.

In the general case in which a properly organised and performed Trial has been statistically analysed, two figures which between them summarise the principal characteristics of the data will emerge—the repeatability  $r$  and the reproducibility  $R$ . By definition<sup>8</sup>, these terms mean that if duplicate analyses of the same sample are made by one operator in any one of the participating laboratories, on 19 out of 20 occasions the difference between the results will not be greater than  $r$ , whereas if the two analyses are made in different laboratories the difference will not be greater than  $R$ . In probability language, the risk of a greater difference will be 5 per cent. If a higher degree of confidence is required, the corresponding figures for a 1 per cent. risk are 1.4 times  $r$  or  $R$ ; for a 0.1 per cent. risk the factor is 1.7.

Armed with this knowledge, the Public Analyst can work out for himself how to fix his Criterion. For example, an actual Trial<sup>4</sup> involved 18 laboratories that determined total fat in cocoa powder. The results from three laboratories were excluded on statistical grounds; those from the remaining 15 gave a mean of 15.95 per cent. and the values of  $r$  and  $R$  were respectively 0.55 and 1.18 per cent. The quantities  $r$  and  $R$ , being estimates of the error within and between the laboratories, are themselves subject to error, and using the knowledge that 15 laboratories were involved their confidence limits can be calculated; for 95 per cent. probability these are 0.4 to 0.9 for  $r$  and 0.9 to 1.6 for  $R$ . Nevertheless, useful deductions can be made from the mean values given above.

The law requires a minimum of 20 per cent. of total fat in the dry matter of cocoa. From  $r$ , the Analyst knows that in his own laboratory, 19 out of 20 duplicate analyses should agree within 0.55 per cent., and that any larger difference needs further investigation, the next obvious step being to have further analyses made. From  $R$  he knows similarly that single analyses in his and another Public Analyst's laboratory will not differ, in 95 per cent. of comparisons, by much more than 1.18 per cent. and in 99 per cent. by much more than  $1.4 \times 1.18 = 1.65$  per cent. If both laboratories make duplicate analyses, the means will differ only in 1 per cent. of cases by more than  $1.65/\sqrt{2} = 1.16$  per cent. (when a sample is analysed  $n$  times in each laboratory,  $R$  should be divided by  $\sqrt{n}$ ). Bearing in mind the "95 per cent. range" of  $R$  as given above, this is of course an approximation; for practical purposes one can say that if duplicate analyses were made in any two of the participating laboratories (and therefore presumably in any other two laboratories comparable in the accuracy and precision of their results) the difference between their respective means will be significantly more than 1 per cent. in only one in a hundred such cases.

Moreover, in about half these unusual occurrences it will not be our Public Analyst's but the other laboratory whose mean result will be the lower, so that the risk of being confronted by a higher result therefrom is only one in two

\* See reference 7 for a discussion of this matter and of the discrepancy between British and American recommendations.

hundred. Again, no Public Analyst is likely to issue an official adverse certificate without having the evidence of at least *three* analyses, so if he adopts a Criterion of, for example, 18 per cent., he will know that no other reputable laboratory is at all likely to report more than 19 per cent. This means that even if a Magistrates' Court decided to accept the evidence of the defence's analyst, a significant deficiency of fat would have been proved. It may be that the average Public Analyst would consider a Criterion of 18 per cent. to be over-cautious and would favour 18.5 per cent. The exact figure to be adopted is a matter for personal judgment; the choice, however, is not a mere guess, as it would be if no Trial had been made, but is based logically on quantitative evidence.

### **Increasing Cost-effectiveness by "Screening"**

It is unfortunately true that to increase the reliability of a result, more work must be done; that all work costs money; and that all Public Analysts have costs in the forefront of their minds. It is logical to devise a routine whereby the less suspicious a sample, the less work is done on it. For some types of determination, all samples can be analysed once and the obviously genuine samples excused duplication. A reasonable routine in the "fat in cocoa powder" case might be—all samples analysed once; all under 20.0 per cent. analysed again (preferably by another operator); all then giving a *mean* result of 19.5 per cent. or less given further examination. In other types of test (e.g. heavy metals) a relatively imprecise but speedy method may be available for separating the obvious "sheep" from the possible "goats", the latter being examined further. All such procedures will much improve the cost-effectiveness of the laboratory.

### **Responsibilities of the Organisers of Trials**

If the object of a Trial is to give participants an opportunity to test their expertise against their peers, then the more that join in the better. But if the object is to test the reliability of a method, then the need for cost-effectiveness requires some restriction on the number of participating laboratories. Two Trials each involving 10 laboratories would enable two methods to be tested with no more expenditure of analytical man-hours than one Trial involving 20, with very little loss of statistical information.

Homogeneity of the test sample before it is divided into sub-samples is vital, as are precautions against changes taking place in transit. It is implied in the design that the "between sub-samples" variance is nil, and any departure could seriously affect the "between laboratories" variance and consequently the value of *R*. Two Trials were recently performed by the same laboratories on different samples of the same food; the improvement in the agreement between laboratories was such as to suggest lack of homogeneity in the sample used in the first Trial.

It is also assumed that the "between methods" variance is nil, and every care must be taken to ensure that this assumption is valid. The method to be used must be specified in every detail, even when it is believed that variations in time, temperature, etc. are not of importance. Failure to realise this has ruined many a Trial in the past. Vagueness must be eschewed. Even the size of the vessel

when following an instruction like "boil vigorously for ten minutes" may be relevant, and what does "vigorously" mean?

Finally (under this sub-heading) the details of the method to be used should, whenever practicable (and sometimes it is not), be circulated with the invitations to participate, so that the laboratory chief can consider carefully what he is being required to do before he accepts the invitation. There may be something in the method—use of a particular instrument, for instance—that would prevent some laboratories from taking part. To know this at once is much better, for the organisers as well as for the laboratory concerned, than having to withdraw after having originally accepted the invitation.

### **Responsibilities of the Participating Laboratories**

The head of an invited laboratory must give very careful consideration to the implications; acceptance is emphatically not a matter to be treated casually. Conscientious collaboration requires adherence to certain rules.

1. The method prescribed must be closely examined to ensure that it is one with which at least some of his analysts are familiar and also that the specification contains no requirement outside the scope of the laboratory. For example, the instruction "ash at  $430 \pm 10^\circ\text{C}$ " would require the use of a thermostatically controlled furnace, and the specification must be adhered to rigidly or not at all.

2. The analyses must be made only by skilled and experienced operators. The results will have to stand comparison with those from the other laboratories; this is a challenge to be accepted or declined, not "fudged".

3. Entry to the Trial must imply the firm intention to carry out the whole of the work specified. Obviously, causes such as illness may make this impossible, and statisticians have methods for compensating for a few missing values; but to drop out completely halfway through the Trial means that the laboratory has completely wasted its own time and in part that of the organisers.

4. All analytical results must be reported to the organisers on completion of the work, without deletion or amendment on the grounds that "it doesn't look right". If more than one operator is taking part they must not be allowed to consult or compare results with each other.

5. "Blanks" must be carried through in full, the method being adhered to rigidly except that there is no sample. This is more important than it may seem; the results can sometimes be enlightening (see later).

All this costs more than using the Trial as an exercise for trainees; but if the job cannot be done properly, it should not be undertaken.

### **The Uses of Collaborative Trials**

Trials have three main uses. First, each provides a searching test of the analytical method employed, and this is often the primary reason for arranging it, the organisers wanting the information before making an official recommendation or incorporating it into legislation. Secondly, it can give the Public Analyst useful help in fulfilling his duties, as we have seen. Thirdly, it enables

each participating laboratory to test its quality against its peers, and this deserves further comment.

The published report of a Trial should, and usually does, contain a table giving all the individual results from every laboratory taking part and a statistical analysis thereof. This begins by noting where it has been found necessary to exclude some results because they were found by a mathematical test to be "outliers", i.e. not consistent with the rest of the data. Occasionally (rarely, but it does happen<sup>4</sup>) all the results from one laboratory are excluded because they show divergence as a whole—usually, because they are all too high or too low. The remaining results are then used to calculate means,  $r$ ,  $R$ , and any other relevant quantities.

The report thus provides all who read it with a picture not only of the value of the method used and the mean results from each sample but also of the expertise of the analysts. Any exclusion of results by the statistician lowers the prestige of the laboratories as a whole. If half or more of the data from one participant are excluded,—much worse, if they all are—this is a cloud the shadow of which falls on all the others. Every laboratory finding that some of its results have been discarded should take this as a reason for internal enquiry and as a warning—"could, and certainly should, do better".

### Another Type of Trial

Sometimes a number of laboratories are asked to make certain determinations by whatever method each normally uses for that kind of analysis. One such Trial<sup>9</sup> asked for the determination of lead and cadmium in six samples of foods. Six Government laboratories, 22 Public Analysts, nine other British laboratories and one West German laboratory took part. The results showed a disturbingly wide range of figures. Taking as an example lead in cabbage, although the results from the six Government laboratories ranged from 0.23 to 0.41 mg/kg. dry weight, with a mean of 0.31, the range of the results as a whole was 0.08 to 0.68, and nine of them were 0.44 to 0.68—unacceptably high if the "official" data are accepted as regards accuracy and precision.

The authors of the paper state that the results taken together show "striking differences between the performances of the laboratories and there is considerable room for improvement". No information is given about the methods used by the laboratories, and the wide spread of the results could be accounted for by:

- (a) the use by some laboratories of a method lacking accuracy and/or precision;
- (b) variations in analytical skill between laboratories;
- (c) unsatisfactory precautions against adventitious contamination by, e.g., dust or incomplete cleaning of vessels;
- (d) insufficient attention to the necessity of careful determination of the "blank".

The last two are well worth consideration; when one is determining a quantity of the order of 0.5 part per million, the extraneous presence of no more than 0.1 p.p.m. represents a 20 per cent. error. Be that as it may, there is an obviously

urgent need for inter-method comparative experiments leading to the establishment of a method of choice accepted by all Public Analysts (and, one would hope, by the Government and other laboratories as well). One wonders how many of the participants in the reported trial had ever previously checked their results in this field of analysis by comparison with any other laboratory on the same sample.

### Conclusion

If properly organised and performed, Trials are a valuable source of external Quality Assurance—a subject much discussed these days<sup>10</sup> as an addition to whatever internal procedures are in use, for they supply objective information that is unobtainable any other way. Just as every laboratory manager must spend money on insurance against such risks as fire and theft, so insurance against the risk of issuing reports based on faulty analyses is not merely desirable but could on occasion even prove crucial.

The author showed the first draft of this paper to J. Markland and learned that the latter had addressed a meeting of the A.P.A. on this subject when he was Public Analyst for Derbyshire. The author is indebted to him for helpful comments and criticisms.

### References

1. Reynolds, S. L., Thorpe, S. A., and Wood, R., *J. Assoc. Publ. Analysts*, 1983, **21**, 47.
2. Jones, D. A., and Wood, R., *J. Assoc. Publ. Analysts*, 1983, **21**, 113.
3. Crimes, A. A., Hitchcock, C. H. S., and Wood, R., *J. Assoc. Publ. Analysts*, 1984, **22**, 59.
4. Judd, H. J., Percival, M. D., and Wood, R., *J. Assoc. Publ. Analysts*, 1984, **22**, 81.
5. Cummings, J. H., Englyst, H. N., and Wood, R., *J. Assoc. Publ. Analysts*, 1985, **23**, 1.
6. Lord, D. W., and Swan, K. J., *J. Assoc. Publ. Analysts*, 1985, **23**, 71.
7. McIntosh, L. J., *J. Assoc. Publ. Analysts*, 1983, **21**, 77.
8. British Standards Institution, "Precision of Test Methods", B.S. 5497, Part 1. London, 1979.
9. Sherlock, J. C., Evans, W. H., Hislop, J., Kay, L. J., Law, R., McWeeny, D. J., Smart, G. A., Topping, G., and Wood, R., *Chem. in Brit.*, 1985, **21**, 1019.
10. Lord, D. W., *Analyt. Proc.*, 1984, **21**, 394.